DOCUMENT RESUME

ED 482 930 TM 035 407

AUTHOR Mislevy, Robert J.; Almond, Russell; Dibello, Lou; Jenkins, Frank; Steinberg, Linda; Yan, Duanli; Senturk, Deniz

TITLE Modeling Conditional Probabilities in Complex Educational Assessments. CSE Technical Report.

INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.; National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.

SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.

REPORT NO CSE-TR-580

PUB DATE 2002-11-00

NOTE 45p.

CONTRACT R305B960002

PUB TYPE Reports - Descriptive (141)

EDRS PRICE EDRS Price MF01/PC02 Plus Postage.

DESCRIPTORS Bayesian Statistics; *Educational Assessment; Item Response Theory; Markov Processes; Models; Monte Carlo Methods; Probability

IDENTIFIERS *Conditional Probability

ABSTRACT

An active area in psychometric research is coordinated task design and statistical analysis built around cognitive models. Compared with classical test theory and item response theory, there is often less information from observed data about the measurement-model parameters. On the other hand, there is more information from the grounding psychological theory, and the task designer's insights into which patterns of skills lead to which patterns of performance. The paper describes a Bayesian approach to modeling these situations, which uses experts' judgments to produce prior distributions for the conditional probabilities in a multivariate latent-variable model, and Monte Carlo Markov Chain estimation to refine the estimated. Task-design schemes and expert judgments are used in the first phase to structure the conditional probability table—that is, conjunctive, compensatory, or disjunctive models, or combinations thereof. Machinery form graded-response item response theory is used to translate experts' judgments about task requirements into prior distributions for model parameters, which in turn imply values for all the conditional probabilities. Bayesian estimation methods are then used to update the distributions for the model parameters in response to observed data. The approach is illustrated with examples from the Biomass biology assessment prototype. (Contains 9 figures, 12 tables, and 25 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made from the original document.

ED 482 930

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
K. Hurst
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

Modeling Conditional Probabilities in Complex Educational Assessments

CSE Technical Report 580

Robert J. Mislevy
CRESST/University of Maryland

Russell Almond, Lou Dibello, Frank Jenkins,
Linda Steinberg, and Duanli Yan
Educational Testing Service

Deniz Senturk
GE

November 2002

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

TM035407

BEST COPY AVAILABLE

Project 3.6 Group Activity on Cognitive Validity
Robert J. Mislevy, Project Director CRESST/University of Maryland

Copyright © 2002 The Regents of the University of California

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

MODELING CONDITIONAL PROBABILITIES IN COMPLEX EDUCATIONAL ASSESSMENTS[1]

Robert J. Mislevy
CRESST/University of Maryland

Russell Almond, Lou Dibello, Frank Jenkins,
Linda Steinberg, and Duanli Yan,
Educational Testing Service

Deniz Senturk
GE

Abstract

An active area in psychometric research is coordinated task design and statistical analysis built around cognitive models. Compared with classical test theory and item response theory, there is often less information from observed data about the measurement-model parameters. On the other hand, there is more information from the grounding psychological theory, and the task designer's insights into which patterns of skills lead to which patterns of performance. We describe a Bayesian approach to modeling these situations, which uses experts' judgments to produce prior distributions for the conditional probabilities in a multivariate latent-variable model, and MCMC estimation to refine the estimates. Task-design schemas and expert judgments are used in the first phase to structure the conditional probability table—that is, conjunctive, compensatory, or disjunctive models, or combinations thereof. Machinery from graded-response IRT is used to translate experts' judgments about task requirements into prior distributions for model parameters, which in turn imply values for all the conditional probabilities. Bayesian estimation methods are then used to update the distributions for the model parameters in response to observed data. The approach is illustrated with examples from the Biomass biology assessment prototype.

---

[1]The Biomass project was supported by Educational Testing Service (ETS) and the College Board from January 2000 through June 2000, and by ETS Research from July through September 2000. Dr. Senturk worked on the project as an ETS summer intern during June and July. Our subject matter expert consultants were invaluable in working through the issues of standards, claims, and evidence that underlie the project, and in offering suggestions along the way for the prototype. They are Ann Kindfield, Dirk Vanderklein, Scott Kight, Cathryn Rubin, Sue Johnson, and Gordon Mendenhall. For providing data on early field trails of Agouti Segment 1, we thank the ETS Summer 2000 Interns, the Weston scholars at Montclair State University and their advisor, Prof. Lynn English, and Russell's buddies at the Knight Dreams comic book shop.

1.0 INTRODUCTION

Insights from cognitive psychology and opportunities from information technology are changing the face of educational assessment today, with new models for what students know and can do, and new ways of capturing data to inform them (National Research Council, 2001). Consequently, an active area in psychometric research is coordinated task design and statistical analysis built around cognitive models (e.g., Adams, Wilson, & Wang, 1997; Embretson, 1985, 1998). Compared with classical test theory and item response theory, there is often less information from observed data about the measurement-model parameters. On the other hand, there is more information from the grounding psychological theory and the task designer's insights into which patterns of skills lead to which patterns of performance. In this paper, we describe a Bayesian approach to modeling these situations, which uses experts' judgments to produce prior distributions for the conditional probabilities in a multivariate latent-variable model, and Markov Chain Monte Carlo (MCMC) estimation to refine the estimates.

Of particular importance is integrating the statistical machinery of Bayesian inference with the substantive issues of task design and evaluation from the very beginning of an application. In the first phase of modeling, task-design schemas and expert judgments are used to structure the conditional probability tables required to model task performance—that is, conjunctive, compensatory, or disjunctive models, or combinations thereof. In the second phase, models from graded-response item response theory (IRT) are used to translate experts' judgments about task requirements into prior distributions for model parameters. In the third phase, Bayesian estimation methods are used to update the distributions for the model parameters in light of observed data.

We illustrate the approach with examples from Biomass, a project carried out at Educational Testing Service (ETS) in 2000. The project produced a computer-based prototype assessment for secondary-school biology, with an emphasis on inquiry skills and model-based reasoning in microevolution and transmission genetics. Four multistage investigative tasks were developed using the "evidence-centered assessment design" approach described in Mislevy, Steinberg, Breyer, Almond, and Johnson (in press). The first segment of one task was pilot tested with 28 summer students at ETS and Montclair State University, and these data will be used to refine the model parameters in the third phase of inference.

2.0 PHASE 0: THE PROBABILITY FRAMEWORK

This section describes the structure and notation we will use for modeling assessment data and a Bayesian approach to inference in this context.

2.1 A Graphical Model for Assessment

For each Student i, let $S_i \equiv (S_{i,1}, \ldots, S_{i,N})$ be a collection of variables characterizing that student's knowledge, skills, or abilities in some domain of interest. We refer to this set of variables and a joint probability distribution as a *student model*. At any point in time, we represent our knowledge about that student's proficiency by a probability distribution. The prior $\Pr(S_i)$ is usually based on the distribution of these skills in the population of interest. We are interested in $\Pr(S_i \mid \mathbf{X}_i)$, where $\mathbf{X}_i = \{X_{i1}, \ldots, X_{iM}\}$ are observations from the student's responses to a collection of M tasks (Almond & Mislevy, 1999). A task may yield more than one observation, as when multiple aspects of a complex performance are evaluated or several questions are asked about the same stimulus materials. In this case, X_{im} is vector-valued, and observations within Task m will be denoted X_{ijm} with j indexing observations within Task m.

If we knew $\Pr(\mathbf{X}_i \mid S_i)$, we could apply Bayes theorem to calculate $\Pr(S_i \mid \mathbf{X}_i)$. Usually we assume that the observations from different tasks are conditionally independent given the student model variables. Thus we consider *evidence models* $\Pr(X_{im} \mid S_i)$, in which the observable(s) in Task m, or X_{im}, is (are) conditionally independent of all observable variables of other tasks and all but a subset of student-model variables. In particular, $\Pr(X_{im} \mid S_i) = \Pr\left(X_{im} \mid S_i^{(m)}\right)$, where $S_i^{(m)} \subset S_i$. We call $S_i^{(m)}$ the *footprint* of evidence model m. Section 3 notes the advantages of defining re-usable evidence structures and conformable task schemas, where the relationships between student-model variables and evidence model variables have been worked out and can be used as skeletons for creating many individual tasks.

At the heart of evidence model m is a collection of conditional probability tables, one for each variable in the evidence model. The case we will address is the multivariate latent class model, in which all student model variables S_i and all the observable variables $X_{i,m}$ are discrete. We may therefore represent the distribution $\Pr(S_i)$ as a discrete Bayesian inference network (Jensen, 1996), as well as joint distributions of the form

$$\Pr(S_i, X_{i1}, \ldots, X_{iM}) = \Pr(S_i) \prod_{m=1}^{M} \Pr\left(X_{im} \mid S_i^{(m)}\right). \quad (1)$$

We will refer to $\Pr(S_i)$ as an SM-BIN fragment and $\Pr(X_{im} \mid S_i^{(m)})$ as an EM-BIN fragment.

Now if we wish to elicit an unstructured prior for $\Pr(X_{im} \mid S_i^{(m)})$, we must specify $| S_i^{(m)} |$ Dirichlet distributions, where $| S_i^{(m)} |$ is the size of the state space of the footprint of Task m. This can be a daunting task. For instance, there are about a hundred observable variables in the Biomass example discussed below, most with three possible values, many with size 18 footprints—over five thousand individual probabilities altogether. In the case of IRT, we have a long history of building evidence models. We aim to draw on that experience to create "structured latent class models" (Formann, 1985) for more ambitious structures for $\Pr(X_{im} \mid S_i^{(m)})$.

2.2 A Bayesian Framework

Gelman, Carlin, Stern, and Rubin (1995, p. 3) describe the first step in Bayesian analysis as setting up a full probability model, or joint probability distribution for all observable and unobservable quantities in a problem. "The model," they continue, "should be consistent with knowledge about the underlying scientific problem and the data collection process." In assessment, scientific knowledge concerns the nature of the targeted knowledge and skill, the ways in which aspects of that knowledge are evidenced in performance, and the features of situations that provide an opportunity to observe those behaviors. The key conditional independence assumption posits that in the main, the aspects of proficiency, expressed in S, account for the associations among responses to different tasks (although we may allow for conditional dependence among multiple responses within the same task); this assumption is manifest in the form of the evidence models described above.

The pertinent variables in assessment obviously include tasks' characteristics and requirements, notably features that have been chosen to elicit observations of the particular kinds, and depending on particular knowledge in ways reflected in the structure of some particular evidence models. Because this presentation focuses on probability-based inference given assessment tasks, we will presume that this design work has been done and the appropriate evidence model structures have been identified. Sections 3 and 4 show how we use the knowledge that the task

authors drew upon when we set prior distributions for the conditional probability tables.

We must thus focus attention on the *X*s, which are potentially observable, and examinees' *S*s, which are not. Structures and parameters that reflect interrelationships among these variables, consistent with our knowledge about them, are also needed. We may start with general forms for the SM-BINs and EM-BINs.

The SM-BIN for Examinee i takes the form of a probability distribution for S_i. An assumption of exchangeability posits a common prior distribution for all examinees before any responses are observed, with beliefs about expected levels and associations among components expressed through the structure of the model and higher level parameters λ; whence, for all examinees i,

$$S_i \sim p(S_i;\lambda).$$

Depending on theory and experience, the distribution for the hyperparameter λ, or $p(\lambda)$, may be vague or precise.

Let $\boldsymbol{\pi}_m$ denote the conditional probabilities in the EM-BIN distributions of Task m. Specifically,

$$\pi_{smjk} = \Pr(X_{imj} = k | S_i = s)$$

that is, the probability of observing a value in response category k for observable variable j of Task m, given that SM variables take the pattern s. Recall that by definition, this probability depends only on $S_i^{(m)}$, the footprint of Task m. The probability of a value x_{imj} as the response of Examinee i for Observable j of Task m is written as

$$p(x_{imj} | \pi_{mj}, s_i^{(m)}),$$

where π_{mj} represents the conditional probabilities for all possible values for Observable j of Task m, given all possible SM patterns.

All tasks using a given EM-BIN structure produce observables in the same forms, furnishing information about the same components of *S*. However, features of the tasks can vary in ways that moderate the relationships. For example, unfamiliar vocabulary and complex sentences tend to make reading comprehension tasks more difficult. One can model πs directly in terms of item features (e.g., Mislevy, Almond, Yan, & Steinberg, 1999). The alternative we will address in this

paper is parametric modeling of the π's, the parameters of which may be informed by expert opinion or empirical data.[2] Denoting all the higher level parameters for Task m by η_m, we write the probability for a given value π_m as

$$p(\pi_m|\eta_m)$$

again with prior knowledge about η_m expressed through higher level distributions $p(\eta_m)$. The complete collection of probabilities for all EM-BINs for all tasks is denoted $\boldsymbol{\pi}$, the parameters for all examinees is denoted **S**, and the responses of all examinees to all tasks is denoted **X**.

The probability model for the responses of N examinees to M tasks can now be written as

$$p(\mathbf{X},\mathbf{S},\pi,\eta,\lambda) = \prod_i \prod_m \prod_j p\left(x_{imj}\middle|s_i^{(m)},\pi_{mj}\right)p(\pi_{mj}|\eta_m)p(\eta_m)p(s_i|\lambda)p(\lambda). \quad (2)$$

Figure 1 represents this model as a generalized form of an acyclic directed graph ("DAG"), with boxes representing repeated elements of the same kind (Spiegelhalter, Thomas, Best, & Gilks, 1995). The structure and the nature of the distributions are tailored to the particulars of an application.

Section 3 concerns the structure of complex assessment tasks and their interrelationship with students' knowledge and skills. Section 4 concerns the way that these considerations can be structured and parameterized in terms of probability distributions, and experts' insights mapped into the formal Bayesian framework. These activities provide the form for (2) in a given assessment context.

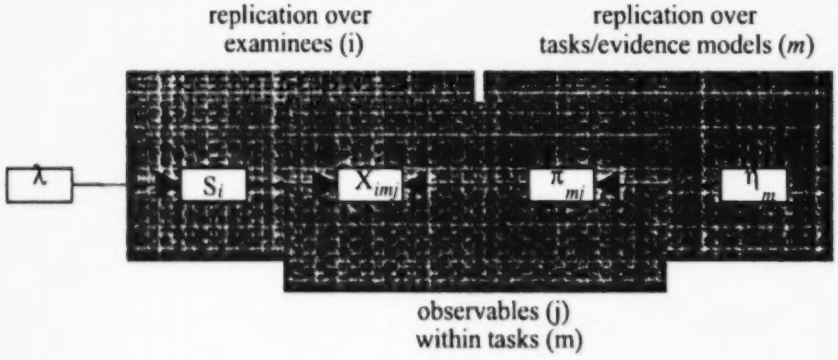

Figure 1. A generic acyclic directed graph of a Bayesian model.

[2]All of these kinds of information may be available, of course, and it is straightforward to incorporate them into a unified model.

Section 5 discusses numerical methods of updating beliefs about examinees and tasks within this framework, as revised posterior distributions for examinee and task parameters, in light of a sample of responses.

3.0 PHASE 1: EVIDENTIARY STRUCTURES BASED ON TASK SCHEMAS

This section sketches the idea of building evidence models around recurring structures among SM variables and observable variables—relationships such as compensatory, conjunctive, disjunctive, and inhibition relationships, and a kind of conditional dependence among observations that is analogous to method factors in factor analysis. Section 3.1 addresses initial implications for Bayesian modeling, and Section 3.2 describes examples from Biomass.

3.1 Recurring Structures in the Evidence Models

Mislevy, Steinberg, and Almond (in press; 2002, see also Mislevy, Steinberg, Breyer, et al., in press) discuss the use of re-usable evidence structures, including the SM- and EM-BINS described above, along with conformable task schemas, as "evidentiary skeletons" around which to create an indefinite number of individual tasks. This is especially advantageous in assessments that use complex tasks, where complexities can include multivariate student models, multivariate observations that depend on the SM variables in different combinations, and dependencies among and within tasks: The structure of situations that elicit valued knowledge and skill can be defined at a higher level of generality, so that the essential relationships among student-model variables and evidence model variables can been worked out and used to build many tasks that may appear quite different on the surface. Following the advice in Mislevy, Steinberg, and Almond (in press, 2002), we want assessment designers to create schemas for creating individual tasks that are built around particular configurations of skills and observations that bear evidence about them. Following the advice in Gelman, et al. (1995), we want psychometricians to incorporate these relationships into Bayesian analyses of observations in these situations.

A test developer who is familiar with a content area and the way students acquire and use knowledge in that area can create situations in which several aspects of skill and knowledge will be required in predictable ways. Some relationships that are familiar from test theory are described below. Section 4 proposes mathematical forms through which they may be expressed.

- *Graded response categories*, as addressed by graded response IRT models. When aspects of a student's performance are evaluated, there may be dimensions of quality that can be described as a sequence of increasingly valued equivalence classes. Performances rated in higher categories are more likely from students with more of whatever combinations of skills are required in the task, while performances in lower categories are more likely from students with less of that proficiency.

- *Conjunctive relationships*, as in binary skills latent class models (Haertel, 1984). Multiple skills are required for performance, and lacking any of them causes lower levels of expected performance. These relationships correspond to AND-gates in logic.

- *Disjunctive relationships*, which correspond to OR-gates. Multiple skills are required, and increasing values of any of them causes higher levels of expected performance.

- *Compensatory relationships*, as in multiple factor analysis (Thurstone, 1947). Multiple aspects of skill or knowledge are involved in performance as captured in an observable variable, and higher levels of those skills imply increasing probabilities of higher levels of the outcome.

- *Inhibition relationships*, (or, stated positively, "enabling relationships") as when a modicum of reading skill is needed to read the directions for more challenging listening tasks in language assessment (Hansen, Forer, & Lee, 2001). Multiple skills are required, but only relatively low values for the "inhibitor" variables. Once these requirements are met, level of performance depends mainly on the other variables.

- *Conditional dependence*, as found among ratings of different aspects of the same performance and among items that share common stimulus materials (Wainer & Kiely, 1987). Conditional dependence concerns relationships among multiple observable variables, indicating that they are related in ways beyond those implied by just the SM variables in their footprint. Ignoring these dependencies results in "double-counting" the information they provide. They are handled in factor analysis with so-called method factors, on which only the affected variables have loadings. Analogous approaches have been implemented in IRT by Bradlow, Wainer, and Wang (1999) and Gibbons and Hedeker (1992).

The utility of these basic structures can be extended by chaining, catenating, or layering them in order to model more complex relationships. Although these relationships can be estimated from data, substantive considerations and design practice can provide strong prior knowledge about the structure of any given task.

3.2 Biomass Examples

As mentioned above, the prototype assessment developed in Biomass addressed inquiry skills and model-based reasoning in the context of microevolution and transmission genetics. Figure 2 is the full Biomass student model **S**, which consists of 15 variables. Each variable has been defined as having three ordered levels of proficiency: High, Medium, and Low (H, M, L). The ovals represent the SM variables; the squares represent probability distributions; and the edges represent the dependence relationships among variables. Their forms will be discussed in Section 5. The variables each concern some aspect of disciplinary knowledge (DK; e.g., the Mendelian model, denoted *DKMendel*), working knowledge (WK; e.g., taking steps in the inquiry process using relevant disciplinary knowledge, denoted *WKInqry*), or integrated knowledge (IK; e.g., reasoning through models across systems or levels of organization, denoted *IKSysOrg*). The model depicts the hierarchical organization of disciplinary, working, and integrated knowledge that was indicted by both our subject matter consultants and standards documents from the domain (e.g., American Association for the Advancement of Science, 1994).

Four multistage investigative tasks were developed, each consisting of a sequence of segments that a student would work through in the course of the larger task. Each segment presented information about results from any previous segments that were needed in the current segment, in order to reduce dependencies across segments. As the examples below illustrate, however, dependencies did occur within segments. A total of 48 evidence models were needed to manage incoming information about students' proficiencies, with several EM-BIN structures appearing more than once. Each EM-BIN fragment contained between 1 and 10 observable variables, and had from 1 to 4 student-model variables in its footprint.

The examples we will address are from the first segment of an investigation in transmission genetics, which we call "Agouti1." The student José discovers a population of mice, notes how many mice have each of four coat colorings, and decides to investigate the mode of inheritance of coat color in mice. This segment yields 14 observable variables, each providing a single categorical response on a 3-point scale. These variables provide evidence about *DKMendel* and/or *WKInqry* through four EM-BINs organized around clusters of related observations:

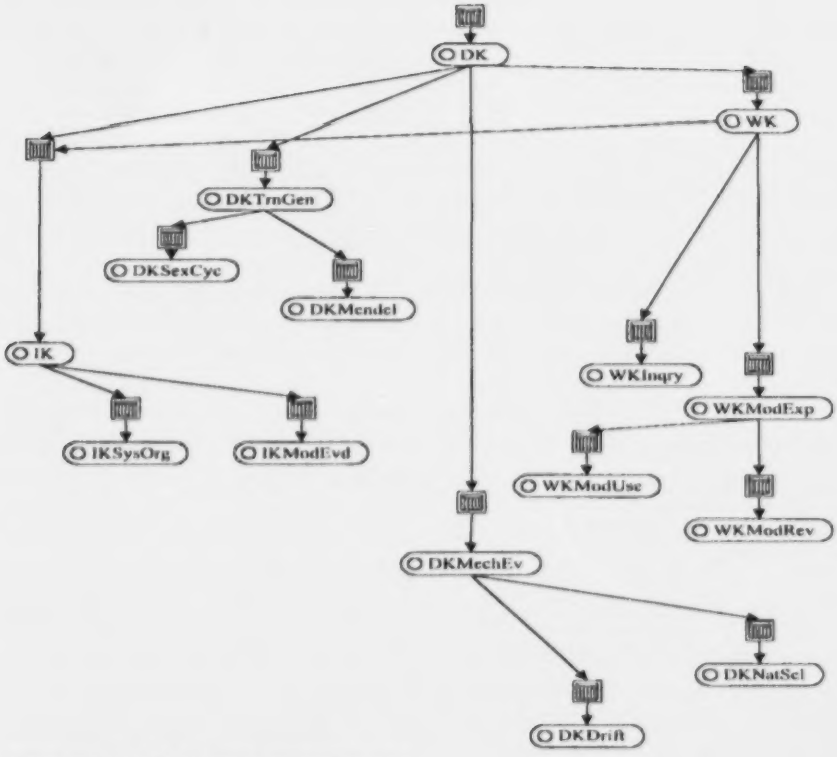

Figure 2. The full Biomass student model.

- *Evidence Model 1* (EM1) concerns aspects of a student's diagrammatic expression of José's verbally stated hypothesis about the mode of inheritance. Figures 3 and 4 show a similar diagram and hypothesis, illustrating how a student would drag and drop elements from a palette of symbols and terms to express her hypothesis. Some of the observable variables concerned the degree of correctness of the elements in given drop targets; for example, on a 1-3 scale, how accurately the dominance relation the student constructed matched Jose's working hypothesis. Others concerned the consistency among different portions of the constructed response.

- For example, Jose posited a dominance relationship, but if a student indicated co-dominance and genotype/phenotype combinations that were consistent with co-dominance, then the observable variable concerning consistency between mode of inheritance and expression of characteristics received a high value. Seven distinct aspects of this solution are captured as values of observable variables, all providing evidence about *DKMendel* but probably dependent beyond their relationship through that SM variable.

Here are the results of José's crosses:

Cross	agouti	agouti-tan	black-tan
agouti ♀ x agouti ♂	11 (69.5♂)		
agouti-tan ♀ x agouti-tan ♂	3 (29.1♂)	7 (39.4♂)	2 (19.1♂)
black-tan ♀ x black-tan ♂			10 (59.5♂)

Based on these results, José thinks that:

- the gene for coat color is in an autosome,
- there are two alleles for this gene in the population, and
- when the two alleles are in the same individual, they both show up in that individual's coat color.

This is José's hypothesis about the mode of inheritance of this gene for coat color in mice.

In order to formalize José's hypothesis, drag symbol(s) or phrase(s) from the toolbox at left to the appropriate columns. Use symbols to complete phrases you have chosen.

Toolbox	Chromosome Type	Alleles	Dominance Relationships	Possible Phenotypes' Corresponding Genotypes
Ag-1 ag-1				/
Ag-2 ag-2				
K An Y				/
...is dominant with respect to...				/
...is recessive with respect to...				
...is co-dominant with respect to...				/
...is incompletely dominant with respect to...				/
				/

Figure 3. A mode of inheritance table, before responses.

BEST COPY AVAILABLE

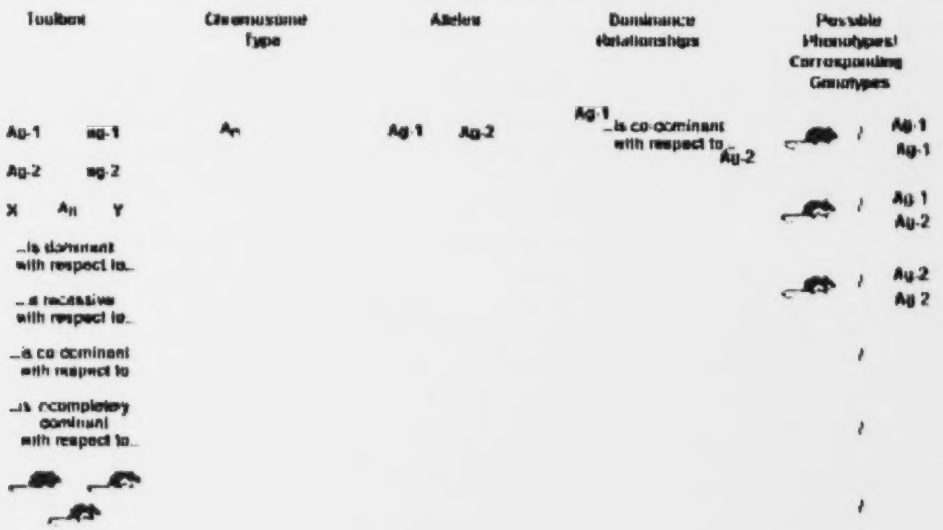

Figure 4. A mode of inheritance table, after responses.

- *Evidence Model 2* (EM2) concerns a table that a student was asked to fill out, in regard to several statements about implications of the mode of inheritance. In each case, the student was to indicate if this statement could be confirmed or rejected on the basis of data from the field population alone, from the offspring of matings of known members of the field population, and from the offspring of matings of the next generation after that. For example, it is a common misconception that if there were more tan mice than black mice in the field population, then tan is the expression of a dominant allele. Maybe, maybe not! There are three variables in this cluster, posited by our experts to depend conjunctively on *DKMendel* and *WKInqry*, and conditionally dependent beyond these joint influences.

- *Evidence Model 3* (EM3) concerns three multiple-choice questions about implications of forms of dominance. *DKMendel* is the only SM variable, and the responses are posited to be conditionally independent.

- *Evidence Model 4* (EM4) asks what José should do next, after having formalized his hypothesis about the mode of inheritance of coat color based on the field population. There is just one observable variable. The key to its solution is a central tenet of inquiry in transmission genetics: Simply generating a hypothesis that is consistent with a field population is not sufficient to conclude a mode of inheritance; one must carry out crosses to test the hypothesis and revise if necessary. Our experts indicated that a

BEST COPY AVAILABLE

student must know at least a bit about the Mendelian model to respond to this question, but the quality of the response would depend mainly on the ability to apply inquiry skills in this domain. The EM-BIN therefore must reflect an *inhibition* relationship, in which a student must be above the Low level of *DKMendel* to have chances at making a high-quality response that increase with increasing levels of *WKInqry*.

4.0 PHASE 2: QUANTITATIVE PRIORS BASED ON EXPERT KNOWLEDGE

This section addresses the conditional probability distributions for observable responses, or the $p\left(x_{imj} \middle| s_i^{(m)}, \pi_{mj}\right)$ and $p\left(\pi_{mj} \middle| \eta_{mj}\right)$ terms. We describe and illustrate the "effective θ" method of assigning conditional probabilities to observable variables that have ordered response categories.

4.1 The Samejima Model for Graded Responses

The most common IRT models are for binary outcomes. The two-parameter logistic model for right/wrong (1/0) responses, for example, is $\text{logit}(\Pr(X_{ij} = 1 \mid \theta)) = a_j(\theta + b_j)$. Samejima's (1969) graded response model extends this model to an observable X_{ij} that can take an integral value from 1 to K. For $k=2,\ldots,K$ define:

$$\Pr(X_{ij} \geq k \mid \theta) = \text{logit}^{-1}(a_j(\theta + b_{jk})), \tag{3}$$

with $\Pr(X_{ij} \geq 1 \mid \theta) = 1$ and $\Pr(X_{ij} \geq K+1 \mid \theta) = 0$. Response category probabilities can be calculated from the differences of equations like (3); for $k=1,\ldots, K$,

$$\Pr(X_{ij} = k \mid \theta) = \Pr(X_{ij} \geq k \mid \theta) - \Pr(X_{ij} \geq k+1 \mid \theta). \tag{4}$$

Figure 5 illustrates response category probabilities for a three-category task, with $a_j = 1$, $b_{j1} = -1$, and $b_{j2} = +1$. For very low values of θ, the lowest level of response is most likely, then as θ increases, probabilities increase for higher valued responses in an orderly manner. A single value of θ specifies the full conditional distribution of all possible response values.

4.2 The "Effective θ" Method

We are interested in finding models for $p\left(x_{imj} \middle| s_i^{(m)}, \pi_{mj}\right)$ in the case where S_i is a discrete Bayesian network; X_{imj} is an discrete variable with ordered states; and π_{mj} are parameters we will specify shortly. We employ the following device. First we pick a fixed set of values for a_{mj} and b_{mj}.[3] Then we define a mapping function $f_{mj}\left(s_i^{(m)}; \pi_{mj}\right)$

[3]Natural extensions for future work are estimating item parameters (being careful to not introduce indeterminacies into the model) and experimenting with different IRT models for multiple-category responses.

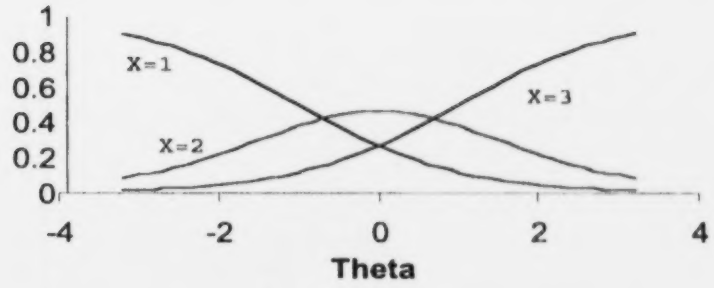

Figure 5. Response category curves from the Samejima graded response model, with a=1 and b=(-1, +1).

to θ_{imj} on $(-\infty,+\infty)$. We can now apply Samejima's graded response model to fill out the tables for Observable j of Task m. Define

$$\Psi_{mjk}(\theta) = \Pr\left(X_{mj} = k \middle| f_{mj}\left(\xi^{(m)}; \pi_{mj}\right) a_{mj}, b_{mj}\right)$$

where the probability is computed with the Samejima graded response model as in (4) with item parameters a_{mj} and b_{mj}.

We gain two advantages with this transformation of the problem. First, in the multivariate case our experts may be comfortable describing the functional form for f_{mj} even if they are uncomfortable with specifying a conditional probability table (e.g., "You have to know how to do A, but then you can solve the problem if you can carry out either procedure B or procedure C"). This is especially true when tasks have been designed from the start around predetermined schemas, for which the structures of recurring evidentiary relationships are already provided.

Second, we have transformed the problem to a scale that is familiar to experts in educational measurement. Thus, they will be more comfortable with the elicitation process on this scale. The scale of IRT models is often set by standardizing the distribution of θ, and in this metric, a value of -1 for b indicates an item that is somewhat easy for the examinees, 0 a typical item, and +1 a somewhat difficult item; further, a parameters typically range from about .3 to 3. When the expert says she expects an item to be easy for the intended population, or that responses will be fairly strongly related to proficiency, we have a good idea of what the a and b parameters will be. If we are planning to refine the evidence models with pretest

data, we can elicit initial opinions in the form of verbal parameters (e.g., "hard" or "easy") that are assigned to numerical priors predefined by psychometricians.

We describe this setup for the one-dimensional and multidimensional cases below and then show how the same approach can also be used to relax the assumption of independent observations.

4.3 EMs With a Univariate Footprint

4.3.1 Basic Formulas

We begin with the case in which an observable X_{mj} has only one SM parent, which we will denote $S^{(m)}$. We define the conditional probabilities $p_{smjk} = p(X_{mj}=k \mid s^{(m)})$ using the projection function $g_{mj}(\cdot)$, a monotonic function of the levels of $S^{(m)}$, which we then enter into a Samejima graded response model with fixed item parameters a_{mj} and b_{mj}. (In particular, we fix all a_{mj}s at 1.) Assuming the levels of $S^{(m)}$ are roughly equally spaced and coding H=1, M=0, and L=-1, a linear function on the index i, $g_{mj}(i) = c_{mj} i + d_{mj}$ gives us just two parameters to elicit from an expert no matter how many states of $S^{(m)}$ or X_{mj} there are. We interpret θ_{mj} as a student's proficiency specific to whatever aspect of performance is captured by Observable Variable j of Task m, and $g_{mj}(\cdot)$ as the projection of $S^{(m)}$ into that space. The constant parameter d_{mj} is related to the average difficulty of the item, and the slope c_{mj} depends on the ability of the task to discriminate among levels of $g_{mj}(S^{(m)})$.

We have thus far specified a structure for terms of the form $p\left(x_{imj} \middle| s_i^{(m)}, \pi_{mj}\right)$, where the hyperparameters π_{mj} specialize to c_{mj} and d_{mj}. We may now suggest forms for the $p(\pi_{mj})$ terms, or more specifically, $p(c_{mj}, d_{mj})$. Leaning on intuition from IRT, we propose for c_{mj} a truncated normal distribution—a N(1,1) distribution, left truncated at 0—and for d_{mj}, a normal distribution with a variance of 1 and a mean γ_{mj} based on expert opinion:

$$\gamma_{mj} = \begin{cases} -1 & \text{for a rather hard item} \\ -.5 & \text{for a harder than typical item} \\ 0 & \text{for a typical item} \\ +.5 & \text{for an easier than typical item} \\ +1 & \text{for a rather easy item} \end{cases}$$

Thus,

$$p(c_{mj}, d_{mj}) = p(c_{mj})p(d_{mj})$$
$$= N^{+}(1,1)N(\gamma_{mj},1)$$

where $N(\mu,\sigma)$ represents the standard normal density with mean μ and standard deviation σ, and $N^{+}(\mu,\sigma)$ is a normal density restricted to $(0,+\infty)$.

4.3.2 An Example From Biomass

Evidence Model 3 concerns three conditionally independent responses to multiple-choice items, modeled as depending on *DKMendel* only. Figure 6 depicts the EM-BIN as an acyclic directed graph. Each item has three ordered possible outcomes, which correspond to a correct response, plausible distractors, and implausible distractors. Our experts said all three are items of typical difficulty, so the initial conditional probability tables will be the same for all three items. Centering the indices for *DKMendel* at –1, 0, and 1 for convenience, we define $g_{3j}(i) = c_{3j}i + d_{3j}$ and set $\gamma_{mj} = 0$. We will begin MCMC estimation with starting values of 1 and 0 for c_{3j} and d_{3j}, so the states (L, M, H) will be mapped to θ_{3j} values of -1, 0, and +1 respectively. The item parameters $a = 1$ and $b = (-.5,+.5)$ are used for the graded response IRT structure into which θ_{3j} is mapped. Table 1 gives conditional response probabilities that correspond to our initial values for *c* and *d*.

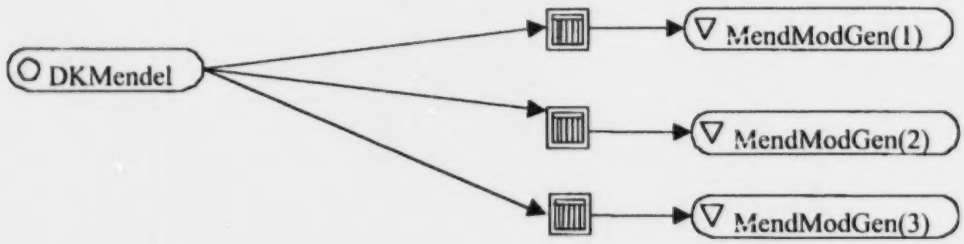

Figure 6. A directed acyclic graph for Evidence Model 3.

4.4 EMs With Multivariate Footprints

Now suppose that $S^{(mj)} = (S_1^{(m)},\ldots,S_L^{(m)})$. It is necessary to construct a projection function $f_{mj}(S_i^{(m)})$ from a vector of SM variables. Before describing some projections that are appropriate for some common evidentiary relationships, we mention three categories into which they may be classified. The first two are elaborations of the linear mappings discussed above for the univariate case.

Table 1

Initial Conditional Probability Distributions for All Three Observables of Task 3

DKMendel		Pr($X = k$)		
Index[a]	q[b]	Low	Medium	High
-1	-1.00	0.62	0.20	0.18
0	0.00	0.38	0.24	0.38
1	1.00	0.18	0.20	0.62

[a]Low = -1, Medium = 0, High = 1.

[b]q = 1.00*index + 0.00.

- *Combinations of linear mappings.* For $l = 1,...,L$, first define a linear mapping $g_{mjl}(i_l) = c_{mj} i_l + d_{mj} \equiv \theta_{imjl}$ that specifies the marginal influence of $S_l^{(m)}$ as to performance for Observable j on Task m. Then define a function $h_{mj}(\theta_{imj1},\ldots,\theta_{imjL})$ that describes how the skills interact to produce proficiency for this particular outcome. The compensatory and inhibitor functions in the following discussion take this form. As in the univariate case, if we assume the skill levels are roughly equally spaced, we can describe that relationship with two parameters per skill.

- *Linear mappings of combinations.* First define a function $r_{mj}(i_1,\ldots,i_L) = i^*_{smj}$ of the indices of L SM variables that describes the structure of their required interaction, such as a maximum or a minimum. Then define a linear mapping $t_{mj}(i^*_{Smj}) = c_{mj} i^*_{Smj} + d_{mj} \equiv \theta_{imj}$ that adjusts for overall difficulty and sensitivity. The conjunctive and disjunctive functions discussed below take this form. Only two parameters are required in this case.

- *Everything else.* Many other structures mapping from multivariate skills to a univariate effective θ can be constructed as the need arises, such as leaky conjunctions and disjunctions, and logical exclusions and necessities.

It is also possible to construct chains of these combining functions, as we shall do with Evidence Model 2 in the Biomass example.

4.4.1 Compensatory Relationships

The most common function for modeling compensatory relationships is weighted sums or averages, as in multiple factor analysis (Thurstone, 1947). We can describe two variations on this theme to use with the effective θ method.

The first is simply the sum of linear mappings for each SM variable involved. That is, for $l = 1,...,L$, $g_{mjl}(i_l) = c_{mjl} i_l + d_{mj} \equiv \theta_{imjl}$, then $h_{mj}(\theta_{imj1},...,\theta_{imjL}) = \sum_l \theta_{imjl}$. The advantage of this formulation is that the relevance and difficulty of some aspect of performance can be assessed with respect to each of the requisite skills, and information about these factors may be available from experts and/or task features.

The individual difficulties are not well determined by response data, as seen by rewriting h_{mj} as $\sum_l c_{mjl} i_l + \sum_l d_{mj}$. The latter sum is tantamount to the item difficulty parameter that is used in compensatory multivariate IRT models (e.g., Reckase, 1985); component-wise difficulties are not identified without additional structures across items. This form may be preferred if information about task features induces informative priors about the d_{mj}s. An alternative formulation is $f_{mj}(s^{(m)}) = \sum_l c_{mjl} i_l + int_{mj}$, where int_{mj} is a single intercept parameter for Item j of Task m.

We postpone illustrating a compensatory relationship until the following section, since the set of Agoutil EMs does not include a simple compensatory relationship but does have conditional dependence relationships that are handled in a very similar way.

4.4.2 Conditional Dependence

Standard IRT models presume that all observable variables are independent given student proficiency. This assumption breaks down for tasks that yield multiple observations, because all can be affected by familiarity with the topic, previous exposure, misunderstandings of the setup, or transitory distractions. We can model this situation by introducing into the evidence model an independent *context* skill variable to allow for relationships among observables within Task m. *Context*, which we may denote by C_m, is then treated as an extra parent of all the observations j within Task m. Other than being discrete rather than continuous, this is how conditional dependence was handled by Bradlow, Wainer, and Wang (1999) in IRT and by Gibbons and Hedeker (1992) in the factor analysis of binary variables.

4.4.2.1 Basic Formulas

Let $S^{(m)}$ be the footprint of Task m, which provides J observables X_{m1}, ..., X_{mJ}. If $f_{mj}\left(S_i^{(m)}\right) \equiv \theta_{imj}$ would be the effective θ for calculating conditional probabilities for Observable mj under conditional independence, we define

$$\theta^{*}_{imj} \equiv f^{*}_{imj}\left(S_i^{(m)}, C_m\right) \equiv \theta_{imj} + e_{mj} i_{Cm},$$

where i_{Cm} is the index of the context variable for Task m (centered around zero for convenience) and e_{mj} is the strength of the intratask dependence as it applies to Observable j. Conditional independence obtains when $e_{mj} = 0$.

The rationale is easiest to see when Context takes only two values, which can be coded as –1 and +1 without loss of generality. A set of θ_{imj}'s would map values of the SM variables $S_i^{(m)}$ into conditional probabilities independently for each Observable mj. But now there are two sets of θ^{*}_{imj} 's, one in which all values are higher than their corresponding θ_{imj}'s by appropriate e_{mj}'s and another in which all are lower by the same e_{mj}'s. An examinee is characterized by an unknown value i_{Cm} that determines which of these two (off)sets actually applies to that examinee. It is marginalizing over the possible i_{Cm} values, when the same one applies to all observables in Task m, that affects conditional dependence.

4.4.2.2 An Example From Biomass

Evidence Model 1 concerns the "mode of inheritance" (MOI) table, which yields seven observable variables. Each is posited to depend on only one SM variable, *DKMendel*, but all are allowed to be conditionally dependent beyond that. Therefore, a context variable is introduced pertaining to all the observables extracted as evaluations of distinct aspects of this same solution to this complex task. Figure 7 depicts this structure. Note that no distribution is shown for *DKMendel*; it is only a "stub" in the EM-BIN fragment. A distribution is included for the Context variable, however. It is local to this task only. We define the Context for the mode-of-inheritance task (abbreviated C_{EM1}) to have two values, High and Low, which we code as –1 and +1 respectively. Again, each observable has three possible outcomes, which correspond to High, Medium, and Low responses (e.g., correct, partially correct, incorrect; or correct, incorrect but consistent, and incorrect and inconsistent).

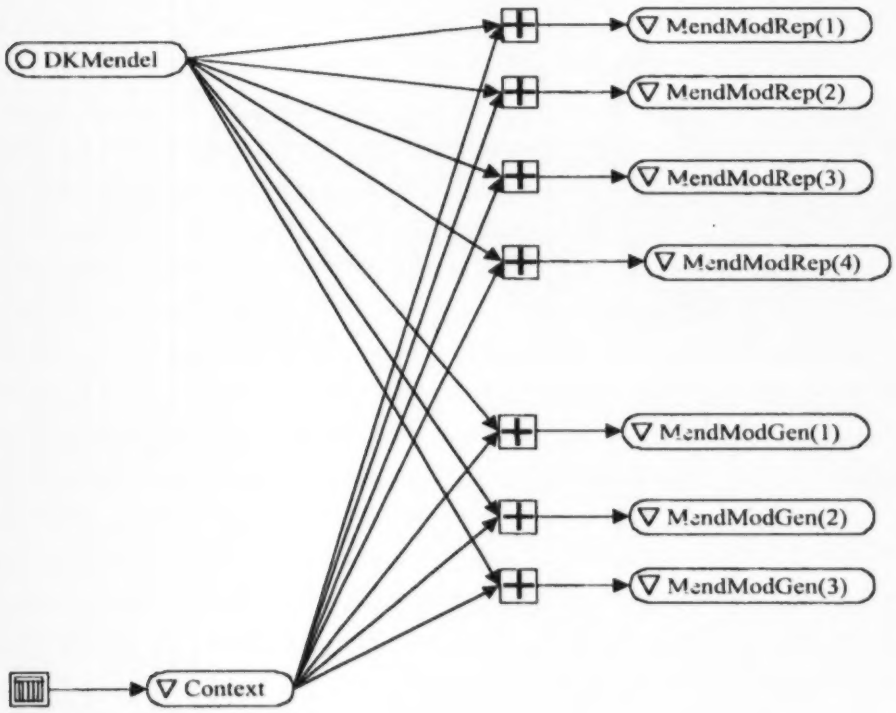

Figure 7. A directed acyclic graph for Evidence Model 1.

Table 2 is a table of initial conditional probability distributions for Observable 1 of Task 1, which our experts identified as easier than typical. They were calculated as follows:

- $\theta_{11} = c_{11} i_{DKM} + d_{11}$
- $\theta_{11}^{*} = \theta_{11} + e_{11} i_{C_{EM1}}$
- Initial values: $c_{11} = 1$, $d_{11} = +1$, $e_{11} = .5$

4.4.3 Conjunctive and Disjunctive Relationships

4.4.3.1 Basic Formulas

Simple conjunctive and disjunctive relationships can be structured as logical operations on the values of SM variables, then mapped linearly to the effective scale. A conjunctive model posits that all skills in a set are required, and the lowest of

Table 2

Initial Conditional Probability Distributions for Observable 1 of Task 1

DKMendel		*Context*		Pr($X = k$)		
Index[a]	q_{11}[b]	Index[c]	q_{11}^*[d]	Low	Medium	High
-1	0.00	-1	-0.50	0.50	0.23	0.27
-1	0.00	1	0.50	0.27	0.23	0.50
0	1.00	-1	0.50	0.27	0.23	0.50
0	1.00	1	1.50	0.12	0.15	0.73
1	2.00	-1	1.50	0.12	0.15	0.73
1	2.00	1	2.50	0.05	0.07	0.88

[a]Low = -1, Medium = 0, High = 1. [b]$\theta_{11} = c_{11} i_{DKM} + d_{11} = 1.00\, i_{DKM} + 1.00$.

[c]Low = -1, High = 1. [d]$\theta_{11}^* = \theta_{11} + e_{11} i_{C_{EM1}} = q_{11} + .5\, i_{EM1}$.

them determines the possibilities of performance. If *DKMendel* and *WKInqry* combine conjunctively to produce a response, for example, a student who is High on *DKMendel* and Low on *WKInqry* is Low on the conjunction. For a conjunctive relationship, then,

$$r_{mj}(i_1,\ldots,i_L) = \min_{\ell}(i_1,\ldots,i_L).$$

A disjunctive model posits that there are several skills that could be used to solve a problem, regardless of the status of others, so it is the highest of them that determines performance. If *DKMendel* and *WKInqry* combine disjunctively to produce a response, the same student who is High on *DKMendel* and Low on *WKInqry* is High on the disjunction. For a disjunctive relationship,

$$r_{mj}(i_1,\ldots,i_L) = \max_{\ell}(i_1,\ldots,i_L).$$

In either case, the logical function can be followed by a linear rescaling with parameters *c* and *d*.

4.4.3.2 An Example From Biomass

Figure 8 shows the structure of Evidence Model 2. Note the chaining, with the conjunction of *DKMendel* and *WKInqry* followed by a compensatory combination with the Context variable C_{EM2} for this set of three observable variables. (Note that this is a different variable from the context variable for the mode-of-inheritance table discussed above.) Tables 3 and 4 show the construction of initial conditional probabilities for the first observable in this evidence model, which our experts

expected to be a little harder than usual. Two tables are used to highlight the conjunctive mapping. They were calculated as follows:

- $\theta_{21} \equiv r_{21}(i_{DKM}, i_{WKI}) = \min(i_{DKM}, i_{WKI})$, with {Low, Medium, High} coded {-1,0,1} for both variables.
- $\theta^{\bullet}_{21} = c_{21}\theta_{21} + d_{21}$ (shown as Table 3 with initial values c_{21} = 1 and d_{21} = -.5)
- $\theta^{\bullet\bullet}_{21} = \theta^{\bullet}_{21} + e_{21} i_{C_{EM2}}$ (shown as Table 4 with initial value e_{21} = .5)

4.4.4 An Inhibition Relationship

In a simple inhibition (or enabler) relationship, one variable must attain a minimal value in order for another variable's values to produce an effect. There is a hurdle that must be overcome.

4.4.4.1 Basic Formulas

Consider an observable variable X_{mj} with a multivariate footprint $S^{(m)}$, such that $S_1^{(m)}$ inhibits the relationship between X_{mj} and its remaining SM parents $\left[S_2^{(m)},\ldots,S_M^{(m)}\right]$. Denote by $\theta^{-}_{mj} \ldots f^{-}_{mj}\left(S^{(m)}\right)$ the mapping from $\left[S_2^{(m)},\ldots,S_M^{(m)}\right]$ to

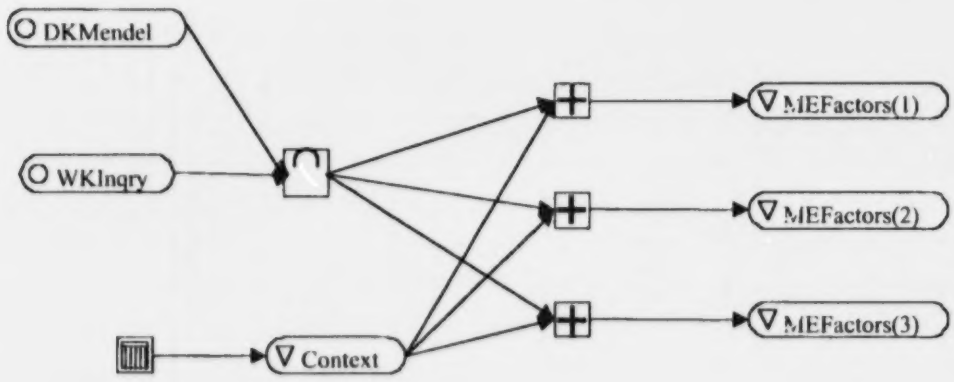

Figure 8. An acyclic directed graph for Evidence Model 2, showing the conjunction of *DKMendel* and *WKInqry*, followed by a compensatory relationship with a *Context* variable that introduces conditional dependence.

Table 3

Effective q for a Conjunction of *DKMendel* and *WKInqry*

DKMendel Index[a]	*WKInqry* Index[a]	Minimum (*DKM*,*WKI*)	q^b
-1	-1	-1	-1.50
-1	0	-1	-1.50
-1	1	-1	-1.50
0	-1	-1	-1.50
0	0	0	-0.50
0	1	0	-0.50
1	-1	-1	-1.50
1	0	0	-0.50
1	1	1	0.50

[a]Low = -1, Medium = 0, High = 1. [b]$q = 1.00*\min(i_{DKM}, i_{WKI}) + -0.50$.

Table 4

Initial Conditional Probabilities Resulting From a Compensatory Relationship Between a *Context* Variable and the Conjunction of *Dkmendel* and *Wkinqry*

Conjunction	*Context*		Pr($X = k$)		
q^a	Index[b]	q^{*c}	Low	Medium	High
-1.50	-1	-2.00	0.82	0.11	0.08
-1.50	1	-1.00	0.62	0.20	0.18
-0.50	-1	-1.00	0.62	0.20	0.18
-0.50	1	0.00	0.38	0.24	0.38
0.50	-1	0.00	0.38	0.24	0.38
0.50	1	1.00	0.18	0.20	0.62

[a]$q = 1.00*\min(i_{DKM}, i_{WKI}) + -0.50$. [b]Low = -1, High = 1.

[c]$\theta^* = \theta \quad + e_{11} i_{C_{EM2}} = q + .5\ i_{EM2}$.

effective θ that applies when a student is over the hurdle value i^*—that is, $S_{i1}^{(m)} \geq i^*$—and denote by $\theta_{mj}^{\min} = \min_{S^{(m)}} \left(\theta_{mj}^{-} \left(S^{(m)} \right) \right)$ the minimum value obtained of θ_{mj}^{-}.

The inhibition relationship can be written as

$$\theta_{mj}^{**}\left(S_i^{(m)}\right) = \begin{cases} \theta_{mj}^{-}\left(S_i^{(m)}\right) & \text{if } S_{i1}^{(m)} \geq i^* \\ \theta_{mj}^{\min} & \text{if } S_{i1}^{(m)} < i^* \end{cases}.$$

4.4.4.2 An Example

Figure 9 shows the structure of the EM-BIN fragment for Evidence Model 4, where *DKMendel* is an inhibitor of *WKInqry*—note the stop sign as a symbol for the structure of the distribution. Table 5 gives a set of conditional probabilities that are obtained as follows:

- $f_{\bar{E}M4}(DKMendel, WKInqry) = \theta_4^*(WKInqry) = c_4 i_{WKI} + d_4$, where i_{WKI} is the index of a student's *WKInqry* value, +1, 0, or -1 corresponding to High, Medium, or Low, respectively.
- $\theta_4^{min} = -c_4 + d_4$, the value of θ_4^* obtained when *WKInqry* is in its lowest state, that is, i_{WKI} = -1.
- The hurdle value for *DKMendel* is Medium; that is, i^* = 0 using the same indexing scheme as for *WKInqry*.
- $\theta_4^{**}(DKMendel, WKInqry) = \begin{cases} c_4 i_{WKI} + d_4 & \text{if DKMendel} \geq \text{Medium} \\ -c_4 + d_4 & \text{if DKMendel} = \text{Low} \end{cases}$
- Initial values for c_4 and d_4 are 1 and 0 respectively.

4.5 The Complete Prior Specification for the Biomass Example

This section summarizes the prior distributions we specified for Agouti 1. The focus of the paper is on the EM-BINs, in particular the effective θ mappings

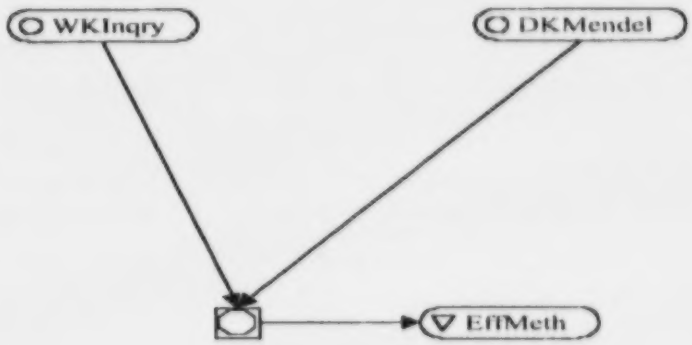

Figure 9. An acyclic directed graph for Evidence Model 4.

Table 5

Initial Conditional Probabilities With *Dkmendel* as an Inhibitor of *WKInqry*

DKMendel Index[a]	*WKInqry* Index[a]	q^{**}[b]	Pr(X = *k*) Low	Medium	High
-1	-1	-1.00	0.62	0.20	0.18
-1	0	-1.00	0.62	0.20	0.18
-1	1	-1.00	0.62	0.20	0.18
0	-1	-1.00	0.62	0.20	0.18
0	0	0.00	0.38	0.24	0.38
0	1	1.00	0.18	0.20	0.62
1	-1	-1.00	0.62	0.20	0.18
1	0	0.00	0.38	0.24	0.38
1	1	1.00	0.18	0.20	0.62

[a]Low = -1, Medium = 0, High = 1.

[b]$$\theta^{**}(DKMendel, WKInqry) = \begin{cases} c_4 i_{WKI} + d_4 & \text{if DKMendel} \geq \text{Medium} \\ -c_4 + d_4 & \text{if DKMendel} = \text{Low} \end{cases}$$

with $c_4 = 1$ and $d_4 = 0$.

discussed in some detail previously—but specifications are required for other parameters as well. We will also give summary statistics for the priors of selected parameters, $DKMendel_i \sim \text{Cat}(\lambda_1, \lambda_2, \lambda_3)$, so that we may compare them with comparable statistics from posterior distributions obtained after the field trial responses.

4.5.1 Priors for Student-Model Variables

In this problem, there are only two student model variables of persistent interest: *DKMendel* and *WKInqry*. However, there are also Context variables, to introduce dependencies among the observables within Evidence Models 1 and 2, that characterize each student. Thus for each Student i, $S_i = (DKMendel_i, WKInqry_i, C_{EM1,i}, C_{EM2,i})$. All are categorical variables, with *DKMendel* and *WKInqry* having three values each (High, Medium, and Low) and C_{EM1} and C_{EM2} having two values each (High and Low). We start with the following prior for *DKMendel*, positing prior exchangeability for students[4]:

$$DKMendel_i \sim \text{Cat}(\lambda_1, \lambda_2, \lambda_3)$$

[4]One could posit different priors for different students in the field trial, based on, say, how many courses they had taken in genetics and how many in science, in general.

where λ_l is the probability that Student i is in State l of *DKMendel*. Dirichlet distributions provide suitable priors. We posit the relatively uninformative prior

$$(\lambda_1, \lambda_2, \lambda_3) \sim \text{Dir}(3,4,3).$$

An intuitive interpretation of this distribution is that it corresponds to the amount of information about the probabilities λ_l that one would have after observing (3-1) + (4-1) + (3-1) = 7 draws, of which 2, 3, and 2 fell into the first, second, and third categories.

The experts anticipate that *DKMendel* and *WKInqry* will be positively associated among students; students with more knowledge about the concepts and representational forms of the Mendelian model will probably have more skill in applying their knowledge. We therefore posit a distribution for *WKInqry* that is conditional on *DKMendel*:

$$WKInqry_i \mid (DKMendel_i = t) \sim \text{Cat}(\lambda_{t1}, \lambda_{t2}, \lambda_{t3})$$

where λ_{tl} is the probability that Student i is in State l of *WKInqry* given that she is in State t of *DKMendel*. We posit for these parameters a set of mild distributions that effect a positive association:

$$(\lambda_{11}, \lambda_{12}, \lambda_{13}) \sim \text{Dir}(5,3,2)$$
$$(\lambda_{21}, \lambda_{22}, \lambda_{23}) \sim \text{Dir}(3,4,3)$$
$$(\lambda_{31}, \lambda_{32}, \lambda_{33}) \sim \text{Dir}(2,3,5)$$

The context variables are posited to be independent of all other SM variables and each other, with

$C_{EM1} \sim$ Bernoulli(.5) and $C_{EM2} \sim$ Bernoulli(.5).

4.5.2 Priors for Evidence Model Parameters

We may organize the remaining prior specifications in terms of evidence models. In all cases, we have used the Samejima graded response model with item parameters $a = 1$ and $b = (0,1)$. We may therefore drop the subscripts indexing items and abbreviate the Samejima model as $\Psi_k(\theta) \equiv \Pr(X = k \mid \theta)$, where the observable and item parameters are apparent from the context of use.

The footprint of EM1 in the student model is $S^{(1)} = (DKMendel, C_{EM1})$. As described in Section 4.4.2.2, EM1 contains seven conditionally dependent observables X_{11} through X_{17}. The conditional probability distributions for these observables have the following form:

$$\Pr\left(X_{i1j} = k \middle| S_i^{(1)}, c_{1j}, d_{1j}, e_{1j}\right) = \Psi\left(\theta^{*}_{i1j}\right)$$
$$= \Psi\left(c_{1j} i_{DKM} + d_{1j} + e_{1j} i_{EM1}\right)$$

$d_{11} \sim N(1,1)$, $d_{1j} \sim N(0,1)$ for j=2,...7, and for j=1,...,7, $c_{1j} \sim N^{+}(1,1)$ and $e_{1j} \sim N^{+}(1,1)$.

The footprint of EM2 is $S^{(2)}$ = (*DKMendel*, WKInqry, C_{EM2}). As described in Section 4.4.3.2, EM2 contains three conditionally dependent observables, X_{21} through X_{23}, which depend on the conjunction of *DKMendel* and *WKInqry*. Thus,

$$\Pr\left(X_{i2j} = k \middle| S_i^{(2)}, c_{2j}, d_{2j}, e_{2j}\right) = \Psi\left(\theta^{*}_{i2j}\right)$$
$$= \Psi\left(c_{2j} \min(i_{DKM}, i_{WKI}) + d_{2j} + e_{2j} i_{EM2}\right)$$

$d_{21} \sim N(-.5,1)$, $d_{22} \sim N(-.5,1)$, $d_{23} \sim N(0,1)$, and for j=1,...,3, $c_{2j} \sim N^{+}(1,1)$ and $e_{2j} \sim N^{+}(1,1)$.

The footprint of EM3 is simply $S^{(3)}$=(*DKMendel*). As described in Section 4.3.2, EM3 contains three conditionally independent observables X_{31} through X_{33}, and

$$\Pr\left(X_{i3j} = k \middle| S_i^{(3)}, c_{3j}, d_{3j}\right) = \Psi\left(\theta_{i3j}\right)$$
$$= \Psi\left(c_{3j} i_{DKM} + d_{3j}\right)$$

and for j=1,...,3, $d_{3j} \sim N(0,1)$ and $c_{3j} \sim N^{+}(1,1)$.

The footprint of EM4 is $S^{(2)}$=(*DKMendel*, *WKInqry*). As described in Section 4.4.4.2, EM4 contains one observable, X_4, which depends mainly on *WKInqry* but is inhibited by *DKMendel*. Thus,

$$\Pr\left(X_{i2j} = k \middle| S_i^{(2)}, c_{2j}, d_{2j}, e_{2j}\right) = \Psi\left(\theta^{**}_{i4}\right)$$

where

$$\theta_4^{**}(DKMendel, WKInqry) = \begin{cases} c_4 i_{WKI} + d_4 & \text{if DKMendel} \geq \text{Medium} \\ -c_4 + d_4 & \text{if DKMendel} = \text{Low} \end{cases},$$

$d_4 \sim N(0,1)$, and $c_4 \sim N^{+}(1,1)$.

4.5.3 Summary Statistics for Selected Parameters

Tables 6, 7, and 8 give summary statistics for the prior distributions described above, along with summary statistics for the posterior distributions that will be described in the following section. The tables concern item parameters, examinee population parameters, and individual examinee distributions respectively. These statistics

Table 6

Summary Statistics of Prior and Posterior Item Parameter Distributions

Evidence model	Parameter groups	Parameter name	Prior mean	Prior *SD*	Posterior mean	Posterior SD	% increase in precision
EM1	Slopes for	c_{11}	1.29	0.79	2.06	0.71	25
	DKMendel	c_{12}	1.29	0.79	1.04	0.63	58
		c_{13}	1.29	0.79	0.95	0.60	72
		c_{14}	1.29	0.79	0.80	0.57	95
		c_{15}	1.29	0.79	0.95	0.61	70
		c_{16}	1.29	0.79	0.79	0.56	101
		c_{17}	1.29	0.79	0.80	0.56	102
	Slopes for	e_{11}	1.29	0.79	1.40	0.55	103
	$Context_{EM1}$	e_{12}	1.29	0.79	2.07	0.53	124
		e_{13}	1.29	0.79	3.18	0.63	58
		e_{14}	1.29	0.79	0.69	0.45	211
		e_{15}	1.29	0.79	3.18	0.63	57
		e_{16}	1.29	0.79	0.70	0.45	209
		e_{17}	1.29	0.79	0.70	0.45	205
	Intercepts	d_{11}	1.00	1.00	2.31	0.60	181
		d_{12}	0.00	1.00	-0.34	0.52	273
		d_{13}	0.00	1.00	-0.26	0.63	157
		d_{14}	0.00	1.00	-2.92	0.59	193
		d_{15}	0.00	1.00	-0.26	0.63	157
		d_{16}	0.00	1.00	-2.91	0.59	192
		d_{17}	0.00	1.00	-2.93	0.59	195
EM2	Slopes for	c_{21}	1.29	0.79	2.09	0.76	8
	conjunction	c_{22}	1.29	0.79	1.11	0.65	46
		c_{23}	1.29	0.79	1.68	0.67	38
	Slopes for	e_{21}	1.29	0.79	0.91	0.60	71
	$Context_{EM2}$	e_{22}	1.29	0.79	0.70	0.48	173
		e_{23}	1.29	0.79	0.65	0.45	213
	Intercepts	d_{21}	-0.50	1.00	-1.55	0.60	178
		d_{22}	-0.50	1.00	-2.94	0.64	150
		d_{23}	0.00	1.00	-0.61	0.50	297
EM3	Slopes for	c_{31}	1.29	0.79	2.37	0.72	22
	DKMendel	c_{32}	1.29	0.79	2.17	0.73	18
		c_{33}	1.29	0.79	2.28	0.72	19
	Intercepts	d_{31}	0.00	1.00	-0.05	0.54	240
		d_{32}	0.00	1.00	1.14	0.51	294
		d_{33}	0.00	1.00	0.14	0.52	271
EM4	Slope	c_4	1.29	0.79	1.03	0.49	161
	intercept	d_4	0.01	1.00	0.81	0.43	433

Table 7

Summary Statistics of Prior and Posterior Population Parameter Distributions

Parameter groups	Parameter name	Prior mean	Prior *SD*	Posterior mean	Posterior *SD*	% increase in precision
	Distribution of *DKMendel*					
	λ_1	0.30	0.14	0.31	0.09	118
	λ_2	0.40	0.15	0.43	0.11	66
	λ_3	0.30	0.14	0.26	0.11	73
	Conditional distribution of *WKInqry* given *DKMendel*					
	λ_{11}	0.50	0.15	0.50	0.15	2
	λ_{12}	0.30	0.14	0.30	0.14	0
	λ_{13}	0.20	0.12	0.20	0.12	-1
	λ_{21}	0.30	0.14	0.31	0.14	-3
	λ_{22}	0.40	0.15	0.40	0.15	-1
	λ_{23}	0.30	0.14	0.29	0.13	7
	λ_{31}	0.20	0.12	0.19	0.11	17
	λ_{32}	0.30	0.14	0.31	0.14	0
	λ_{33}	0.50	0.15	0.50	0.15	8

are based on 50,000 draws from the prior using the Gibbs sampler used in Section 5, but without response data. Using the generic notation of (2), this means drawing from

$$p(\mathbf{S}, \pi, \eta, \lambda) = \prod_i \prod_m \prod_j p(\pi_{mj}|\eta_m)p(\eta_m)p(s_i|\lambda)p(\lambda). \quad (5)$$

Specialized to the Biomass example, S=(*DKMendel*, *WKInqry*), and their higher level parameters generically denoted λ are here parameters in categorical probability distributions. Note that in the Biomass example, the π_{mj} terms are conditional probabilities calculated directly through the Samejima model with effective thetas via task parameters denoted generically by η_m, and are here c's, d's, and e's. This means that the $p(\pi_{mj}|\eta_m)$ are deterministic functions, and the only uncertainty associated with π's is due to uncertainty about η's.

Note that the prior distributions for all the item slopes are identical, whereas the item difficulties vary in accordance with the experts' judgments of their

Table 8

Summary Statistics of Prior and Posterior Student Parameter Distributions

Student	*DKMendel*					*WKInqry*				
	Prior mean	Prior SD	Post. mean	Post. SD	% increase precision	Prior mean	Prior SD	Post. mean	Post. SD	% increase precision
1	2.00	0.77	2.11	0.48	157	2.00	0.82	1.62	0.72	28
2	2.00	0.77	1.03	0.16	2229	2.00	0.82	1.71	0.78	9
3	2.00	0.77	1.01	0.12	4390	2.00	0.82	1.71	0.79	8
4	2.00	0.77	1.63	0.53	114	2.00	0.82	1.67	0.76	16
5	2.00	0.77	2.10	0.47	175	2.00	0.82	1.97	0.80	5
6	2.00	0.77	2.13	0.45	201	2.00	0.82	1.97	0.79	5
7	2.00	0.77	1.83	0.44	206	2.00	0.82	2.15	0.75	18
8	2.00	0.77	1.95	0.50	136	2.00	0.82	1.89	0.73	26
9	2.00	0.77	1.02	0.12	3985	2.00	0.82	1.71	0.78	8
10	2.00	0.77	2.08	0.49	151	2.00	0.82	1.62	0.72	28
11	2.00	0.77	2.14	0.48	163	2.00	0.82	2.00	0.70	36
12	2.00	0.77	1.69	0.51	135	2.00	0.82	1.86	0.74	22
13	2.00	0.77	2.63	0.49	154	2.00	0.82	2.27	0.69	39
14	2.00	0.77	1.66	0.51	133	2.00	0.82	1.91	0.81	2
15	2.00	0.77	2.65	0.48	162	2.00	0.82	2.04	0.70	37
16	2.00	0.77	1.01	0.11	5180	2.00	0.82	1.71	0.78	9
17	2.00	0.77	2.94	0.23	1036	2.00	0.82	2.86	0.37	393
18	2.00	0.77	2.12	0.47	175	2.00	0.82	2.32	0.68	42
19	2.00	0.77	1.75	0.57	84	2.00	0.82	1.64	0.74	20
20	2.00	0.77	1.16	0.37	341	2.00	0.82	1.69	0.77	11
21	2.00	0.77	1.10	0.31	542	2.00	0.82	1.67	0.78	10
22	2.00	0.77	2.88	0.33	453	2.00	0.82	2.83	0.40	322
23	2.00	0.77	2.79	0.41	263	2.00	0.82	2.69	0.50	161
24	2.00	0.77	2.23	0.46	182	2.00	0.82	2.26	0.70	34
25	2.00	0.77	2.63	0.49	152	2.00	0.82	1.94	0.70	35
26	2.00	0.77	1.01	0.10	5490	2.00	0.82	1.70	0.79	8
27	2.00	0.77	2.18	0.44	212	2.00	0.82	2.25	0.71	32
28	2.00	0.77	2.61	0.50	143	2.00	0.82	1.58	0.67	47

difficulties. The values for the examinee priors are means and standard deviations calculated with High = 3, Medium = 2, and Low = 1.

5.0 PHASE 3: REFINEMENT BASED ON FIELD TRIAL DATA

We have spent some effort to build a Bayesian probability framework that expresses our beliefs about the key relationships between knowledge and performance in the Biomass tasks. The probability distributions express the qualitative structure of the relationships, and task and examinee parameters express the quantitative relationships within that structure. We are in a position to update our beliefs with information from some actual observations. The focus of Sections 5.1 and 5.2 is on posterior distributions for task and examinee parameters—refinements of quantitative relationships within the posited structure. More briefly, Section 5.3 discusses criticism of the qualitative structure itself.

5.1 The Markov Chain Monte Carlo Setup

In Bayesian inference, parameters express belief about the nature and magnitude of relationships in observable variables. We are thus interested in posterior distributions for those parameters, which incorporate information from realized observations into our prior beliefs about the structure of the problem. This means conditioning on the particular values of X obtained from students to produce the following posterior distribution, stated first in the generic notation:

$$p(\mathbf{s},\pi,\eta,\lambda|\mathbf{X}) \propto \prod_i \prod_m \prod_j p\left(x_{imj}\middle|s_i^{(m)},\pi_{mj}\right) p(\pi_{mj}|\eta_m)\, p(\eta_m)\, p(s_i|\lambda)\, p(\lambda) \qquad (6)$$

Note the similarity in form between the full Bayesian model for all observations and parameters, given earlier as (2), and the posterior for the parameters, given as (6). The difference is that in (6), the values of the observables are known and fixed. We see that these terms are the difference between the prior distribution for the parameters (5) and their posterior (6).

Monte Carlo Markov Chain (MCMC) techniques provide a general approach to computation in Bayesian inference (e.g., Gelman et al., 1995) that suits the modular construction of assessment we argued for in Section 3. A full treatment of MCMC methods is beyond the current presentation, but the essential idea is to produce draws from a series of distributions that are equivalent in the limit to drawing from the posterior distribution of interest. We used the BUGS computer program (Spiegelhalter et al., 1995) to effect a Gibbs sampling solution in our example. Each iteration produces a value for each parameter in the model, drawn from what is called its "full conditional" distribution: Its distribution is conditional on not only

the data, but a value for every other parameter in the model. In the Gibbs sampler, the values for the other parameters are draws from their full conditional distributions on the previous iteration. Using the general notation and describing the process at the level of blocks of parameters for convenience, the $t+1^{th}$ iteration looks like this:

Draw $\mathbf{S}^{t+1}$ from $p(\mathbf{S}|\pi^t, \eta^t, \lambda^t, \mathbf{X})$;

Draw η^{t+1} from $p(\eta|\mathbf{S}^{t+1}, \pi^t, \lambda^t, \mathbf{X})$;

Calculate π^{t+1} from η^{t+1}; and

Draw λ^{t+1} from $p(\lambda|\mathbf{S}^{t+1}, \pi^{t+1}, \eta^{t+1}, \mathbf{X})$.

Under broad conditions, the distribution of draws from a sequence of iterations converges to draws from a stationary distribution that is the desired posterior, and the empirical distribution of a large number of draws for a given parameter approximates its marginal distribution. Summaries such as posterior means and variances can be calculated (for example, to construct self-contained SM- and EM-BIN fragments).

5.2 Posterior Distributions

Table 9 gives the responses of the 28 students in the field trial, and Tables 6-8, which were introduced at the end of Section 4, give summaries of posterior distributions for parameters conditional on this data. In this section, we offer some observations on these results.

Looking first at the response data, we note immediately a dearth of "2" responses, except for the last observable. In most cases, the students did well or poorly on most aspects of the tasks, without many performances of intermediate quality—even though the average of all the responses, with H = 3, M = 2, L = 1, was 1.62, just about in the middle. The students showed a great range in performance: Nothing better than the lowest response from Students 3 and 26, to a majority of 3s for Student 22.

The items range from very difficult (nobody did better than the lowest response on x_{14}, x_{16}, x_{17}, and x_{22}) to fairly easy (most students answered x_{11} correctly). Did these results accord with the experts' prior expectations? Sort of. There were

Table 9

Observed Responses

Student	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{21}	x_{22}	x_{23}	x_{31}	x_{32}	x_{33}	x_4	Mean
1	3	1	1	1	1	1	1	1	1	1	3	3	1	2	1.50
2	1	1	1	1	1	1	1	1	1	1	1	1	1	3	1.14
3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1.00
4	3	1	1	1	1	1	1	1	1	1	1	1	3	2	1.36
5	3	3	3	1	3	1	1	1	1	1	1	3	3	3	2.00
6	3	1	1	1	1	1	1	1	1	1	3	3	1	3	1.57
7	3	1	1	1	1	1	1	1	1	2	1	3	1	3	1.50
8	3	1	3	1	3	1	1	1	1	2	1	3	3	2	1.86
9	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1.07
10	3	1	1	1	1	1	1	1	1	1	1	3	3	2	1.50
11	3	3	3	1	3	1	1	2	1	1	1	3	3	2	2.00
12	3	3	3	1	3	1	1	1	1	2	1	3	1	2	1.86
13	3	3	3	1	3	1	1	1	1	2	3	3	3	3	2.21
14	3	3	3	1	3	1	1	1	1	1	1	3	1	3	1.86
15	3	1	1	1	1	1	1	2	1	1	3	3	3	2	1.71
16	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1.00
17	3	3	1	1	1	1	1	3	1	3	3	3	3	3	2.14
18	3	1	3	1	3	1	1	1	1	3	3	1	3	3	2.00
19	3	3	1	1	1	1	1	1	1	1	3	1	1	2	1.50
20	1	1	1	1	1	1	1	1	1	1	1	3	1	2	1.21
21	3	3	3	1	3	1	1	1	1	1	1	1	1	1	1.57
22	3	3	3	1	3	1	1	3	1	3	3	3	3	3	2.43
23	3	1	1	1	1	1	1	3	1	2	3	3	3	3	1.93
24	1	1	1	1	1	1	1	1	1	2	3	3	3	3	1.64
25	3	3	3	1	3	1	1	1	1	2	3	3	3	2	2.14
26	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1.00
27	3	1	1	1	1	1	1	1	1	2	3	3	1	3	1.64
28	3	1	1	1	1	1	1	1	1	1	3	3	3	2	1.64
Mean	2.50	1.71	1.71	1.00	1.71	1.00	1.00	1.29	1.00	1.50	1.93	2.36	2.00	2.29	1.64

three items for which they had opinions other than "typical." They expected Observable 1 of Task 1 to be easier, and it turned out to be the easiest one in the study. They expected Observables 1 and 2 of Task 2 to be easier than typical, and they were. But the four observables noted above on which every student was rated Low were not expected to be different from typical. This may be due to the fact that the students in the field trial were not exactly the same as the ones the experts had in

mind as a target population. They thought about how hard a task would be for a student who had been working through a unit on this material and who would be familiar with the notation and expectations used in the prototype. Our field trial students did not have this advantage, which could differ from one task to the next.

The prior distributions we posited for the parameters were fairly mild. Looking at posterior distributions, we see that the information in 14 responses each from 28 students was sufficient to impact distributions for individual students substantially, but it had hardly any effect on belief about the distribution of SM variables (i.e., the λ's).

Task parameter posteriors showed means that departed significantly from their priors. The slopes for Context variables, for example, were initially all at 1.29; posterior means ranged from .65 to 3.18. Intercept means, which were initially at 0 for typical items, ranged from -2.94, for the items on which no one succeeded, to 2.37, for a task on which about two thirds of the students did well. To see the effect of the data on the conditional probabilities, compare Tables 10 and 11 for Observables 3 and 4 of Task 1. They started with the same initial conditional probability tables, since the experts expected both to be about typical. These tables have been calculated through the Samejima structure with the posterior means of their respective task parameters. Note that the revised conditional distributions for Observable 3 show it is much easier than Observable 4, and much more conditionally associated with the other observables within this task.

Table 10

Revised Conditional Probability Table for Observable 3 of Task 1

DKMendel Index[a]	θ_{11}[b]	*Context* Index[c]	θ_{11}^{*}[d]	Pr($X = k$)		
				Low	Medium	High
-1	0.00	-1	-0.50	0.98	0.01	0.01
-1	0.00	1	0.50	0.08	0.11	0.81
0	1.00	-1	0.50	0.95	0.03	0.02
0	1.00	1	1.50	0.03	0.05	0.92
1	2.00	-1	1.50	0.88	0.07	0.05
1	2.00	1	2.50	0.01	0.02	0.97

[a]Low = -1, Medium = 0, High = 1. [b]$\theta_{11} = c_{11} i_{DKM} + d_{11} = 0.95\ i_{DKM} + -.26$.

[c]Low = -1, High = 1. [d]$\theta_{11}^{*} = \theta_{11} + e_{11} i_{C_{EM1}} = \theta_{11} + 3.18\ i_{EM1}$.

Table 11

Revised Conditional Probability Table for Observable 4 of Task 1

DKMendel Index[a]	θ_{11}[b]	*Context* Index[c]	θ_{11}^{*}[d]	Pr($X = k$) Low	Medium	High
-1	0.00	-1	-0.50	0.98	0.01	0.01
-1	0.00	1	0.50	0.93	0.05	0.03
0	1.00	-1	0.50	0.96	0.03	0.02
0	1.00	1	1.50	0.85	0.09	0.06
1	2.00	-1	1.50	0.91	0.06	0.04
1	2.00	1	2.50	0.72	0.16	0.13

[a]Low = -1, Medium = 0, High = 1. [b]$\theta_{11} = c_{11} i_{DKM} + d_{11} = 0.80\, i_{DKM} + 2.92$.

[c]Low = -1, High = 1. [d]$\theta_{11}^{*} = \theta_{11} + e_{11} i_{C_{EM1}} = \theta_{11} + 0.69\, i_{EM1}$.

To quantify the amount of information about the various parameters, the parameter summary tables indicate a percentage increase in precision from priors to posteriors. It is calculated as follows:

$$\%\ \text{Increase in precision} = 100 \times \frac{(\text{posterior SD})^{-2} - (\text{prior SD})^{-2}}{(\text{prior SD})^{-2}}.$$

A value of zero would indicate no new information, and a value of 100 would mean there was twice as much information about a parameter after seeing the data than there was before seeing it.

There are only very modest increases for the parameters of the student distribution—noticeable for the distribution of *DKMendel*, since every student contributes something, with information from all of their responses, but almost none for conditional distributions of *WKInqry* given *DKMendel*. This latter result obtains both because there is less information about *WKInqry* for each student and because the conditional distributions for *WKInqry* would necessarily be based on fewer observations than a marginal distribution for *DKMendel*, even if individual students' values were known with certainty.

Task parameters show increases in precision that are greater than those for student population parameters, but less than for individual students (see next paragraph). In general there are greater increases in precision of intercept parameters than for slope parameters, a finding consistent with experience in IRT. It is intriguing to see that evidence is particularly weak for the slope parameters of the

conjunction of *DKMendel* and *WKInqry* in EM3. Further investigation is needed to determine whether this is a pervasive characteristic of combinations such as conjunctions and disjunctions.

There are substantial increases in precision for the posteriors of individual students, at least as far as *DKMendel* is concerned. These means are calculated as expected values over the coding High = 3, Medium = 2, and Low = 1, so high precision corresponds to probability concentrated on one particular value. Thus, posterior precision is very high for students who performed at high levels on all tasks or at low levels on all tasks; almost all of their posterior probability is on the highest or the lowest value of an SM variable. We learn more about *DKMendel* than about *WKInqry*, mainly because there are more observables that provide information about *DKMendel*. Posterior precision is greater for *DKMendel*, and posterior means can be further from their prior means than is the case for *WKInqry*. With this small field trial data set, we may not learn much about higher level parameters, but even given broad priors that rely on experts' opinions, we are pretty sure that a student who does poorly on most of the observables is Low and a student who does well is High!

5.3 Model Fit

Model criticism is an essential facet of Bayesian (or any other) statistical inference, since the inferences that probability-based reasoning allows us to draw through models are suspect if the data do not accord well with the model. This regrettably brief section outlines the route we are beginning to pursue in examining fit in the kinds of models we have discussed in this paper. The reader is referred to Gelman et al. (1995) and chapters 9-13 of Gilks, Richardson, and Spiegelhalter (1996) for discussions of model criticism in MCMC estimation more generally.

The particular technique we are exploring is the use of *shadow* data sets, created in the course of MCMC iterations. For each observed response x_{imj} in the realized data, we can define another variable y_{imj} that follows exactly the same distribution we have proposed and fit for x_{imj} but is never observed. If our model is correct, the actual data is a plausible draw from the predicted distribution of the shadow data. Thus, the distribution of the shadow data or any summary statistic of it that is accumulated over the MCMC iterations constitutes a tailor-made null distribution against which to evaluate how surprising the data are in light of the model we have proposed. (See Ludlow, 1986, for an example of the usefulness of this approach in

IRT before MCMC techniques were widely available.) Again using the generic notation, we use the following distribution to produce the predictive distribution of the shadow data matrix **Y**:

$$p(\mathbf{Y}|\mathbf{X},\mathbf{S},\P,\ ,\lambda) \propto \prod_i \prod_m \prod_j p\left(y_{imj}\middle|s_i^{(m)},\pi_{mj}\right)p\left(x_{imj}\middle|s_i^{(m)},\pi_{mj}\right)p(\pi_{mj}|\eta_m)p(\eta_m)p(s_i|\lambda)p(\lambda)$$

Table 12 presents one draw of shadow responses. Note that the averages for both observables and students approximate those of the observed data closely, including the observables on which no actual students did better than Low. The lack of "2" responses, except for the final observable, is also replicated. Any statistic of actual responses, such as correlations and joint distributions, as well as the marginal means we have shown, could be calculated on the shadow data set as well. Because the distribution of such statistics could then be accumulated over iterations, an empirical null distribution would be obtained against which to evaluate how typical or how surprising the corresponding feature of the real data was.

One way we used the shadow data was to evaluate an index of examinee fit. Define the fit mean square for Examinee *i* as follows:

$$Z_i = \frac{1}{14}\sum_m \sum_j \left(x_{imj} - E\left(x_{imj}\right)\right)^2 . \qquad (7)$$

where responses are coded H=3, M=2, L=1, and

$$E\left(x_{imj}\right) = \sum_{k=1}^{3} \Pr\left(x_{imj} = k|S_i,\pi_m\right)$$

In iteration *t* of the Gibbs sampler, these quantities can be evaluated conditional on the draws of the task and examinee population parameters. So too can corresponding fit mean squares in which each actual observation x_{imj} in (7) is replaced by its shadow counterpart y_{imj}. The relevant index is the proportion of iterations in which the fit mean square for the *x*'s is greater than the one for the *y*'s. One run with 1,000 iterations produced values across the 28 examinees between .06 for Examinee 25 (the best fit) and .78 for Examinee 18 (the worst fit). Examinee 18's pattern is somewhat uncommon because of High values for the slightly harder-than-typical observables x_{23} and x_{31}, coupled with a low value for the easier-than-typical observable x_{32}. The fact that the highest empirical *p*-value was only .78 caused us some concern about the power of the test. We did a second run with an additional

Table 12

One Set of Shadow Responses

Student	y_{11}	y_{12}	y_{13}	y_{14}	y_{15}	y_{16}	y_{17}	y_{21}	y_{22}	y_{23}	y_{31}	y_{32}	y_{33}	y_4	*y* Mean	*x* Mean
1	3	1	1	1	1	1	1	1	1	1	1	1	2	3	1.36	1.50
2	1	3	1	1	1	1	1	1	1	1	3	1	1	1	1.29	1.14
3	1	1	2	1	1	1	1	1	1	1	2	2	1	3	1.36	1.00
4	3	1	1	1	1	1	1	1	1	1	3	1	1	3	1.43	1.36
5	3	3	3	1	3	1	1	1	1	1	3	3	3	3	2.14	2.00
6	3	1	1	1	1	1	1	1	1	2	3	2	3	3	1.71	1.57
7	2	3	1	1	3	1	1	1	1	1	3	3	2	2	1.79	1.50
8	3	2	3	1	1	1	1	3	1	3	2	2	1	1	1.79	1.86
9	3	1	1	1	1	1	1	1	1	1	1	1	2	1	1.21	1.07
10	3	1	1	1	1	1	1	1	1	1	3	3	3	2	1.64	1.50
11	2	3	2	1	3	2	1	1	1	3	3	3	3	2	2.14	2.00
12	3	3	3	1	3	1	1	1	3	1	3	3	1	2	2.07	1.86
13	3	2	3	1	1	1	1	2	3	3	2	3	3	2	2.14	2.21
14	3	3	3	1	1	1	1	1	1	1	3	1	1	2	1.64	1.86
15	1	1	3	1	1	1	1	1	1	1	3	3	3	1	1.57	1.71
16	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1.14	1.00
17	3	1	2	1	1	1	1	3	1	3	3	3	3	2	2.00	2.14
18	3	3	1	1	1	1	1	3	1	3	3	3	3	3	2.14	2.00
19	3	3	3	1	1	1	1	1	1	1	2	1	1	2	1.57	1.50
20	1	3	1	1	1	1	1	1	1	1	1	1	2	1	1.21	1.21
21	3	1	1	1	1	1	1	1	1	1	1	3	2	1	1.36	1.57
22	3	3	3	1	1	1	1	2	1	2	3	3	2	1	1.93	2.43
23	3	2	1	1	1	1	1	3	2	1	3	3	3	1	1.86	1.93
24	3	1	1	1	1	1	1	1	1	2	3	3	3	1	1.64	1.64
25	3	3	3	1	3	1	1	1	1	3	3	3	2	1	2.07	2.14
26	3	1	1	1	1	1	1	1	1	1	1	2	1	1	1.21	1.00
27	3	1	1	1	1	1	1	1	1	1	1	3	2	3	1.50	1.64
28	3	2	1	1	1	1	2	1	1	2	3	3	2	3	1.86	1.64
y Mean	2.64	1.93	1.75	1.00	1.36	1.04	1.04	1.36	1.18	1.57	2.36	2.29	2.04	1.86	1.67	
x Mean	2.50	1.71	1.71	1.00	1.71	1.00	1.00	1.29	1.00	1.50	1.93	2.36	2.00	2.29		1.64

fictitious response vector, one with High values for the harder observables and Low values for the easier ones:

$$x_{badfit} = (1,1,1,3,1,3,3,3,3,3,3,1,1,1).$$

We were comforted to see that of 20,000 draws of a shadow response pattern to this maximally bad fitting pattern, only 4 had a higher mean square—an empirical *p*-

value of .0002. When a response vector is seriously out of sorts, this index will flag it. (This is just an existence proof, of course; a more serious analysis would run simulations to characterize the specificity and the sensitivity of fit indices constructed in this manner.)

6.0 CONCLUSION: NEXT STEPS

In this paper, we have described an approach to building conditional probability distributions for complex assessments and illustrated the ideas with some specifics we have worked out thus far. There is much to do, along many dimensions.

Substantive issues concern the development of conditional probability model structures that are useful and reusable across applications. As the link between substantive experts' ways of thinking about problems in their domain and statisticians' ways of thinking about parameters and distributions, these structures must both correspond to substantively important aspects of tasks and support sound estimation procedures. We have found this a challenge best met by a small team of experts focused on this goal, whose work provides schemas for complex tasks and skeletons of the evidentiary arguments that underlie them, to be fleshed out as many times and in as many ways as task authors then care to do. The alternative approach of creating complex tasks without considering these issues looks to us like a loser, at least in the context of medium- to large-scale assessment (as we argue in Mislevy, Steinberg, Breyer, et al., in press). The practical benefits of efficiency and reusability are foregone, to be sure, but a more serious loss is the explicit and careful working through of the evidentiary argument. Messick's 1994 paper on performance assessment remains invaluable for thinking about how to design complex tasks. We see our work as fleshing out the psychometric implications of his ideas.

Estimation issues were not the focus of this presentation, but it is clear that attention is required there as well. Obvious steps would be running and monitoring chains of MCMC iterations from multiple starting points and more in-depth investigations of model fit. In particular, the different relationships among SM parents of observables need to be compared. Our experts proposed, and we fit, a conjunctive model for EM3. Would a compensatory model have fit as well, or better? The low efficiency for estimating the item parameters of the conjunctive relationship suggests there may be benefits in a bias toward linearity in models. Other extensions

we have mentioned along the way include more flexible estimation of task conditional probability structures, and incorporation of task features as collateral information about task difficulty parameters. For the Biomass example itself, we should gather and analyze more data, increasing the student sample size and expanding to more tasks.

Operational issues will flow from what we learn in the research described above. That is, what kinds of tools, data structures, interfaces, and building blocks help an organization carry out this work efficiently on a large scale? We have made some progress already in tools for designing task and statistical-model fragments (see the section on the Portal project in Frase et al., 2003). Extensions we see a need for right now include the automatic generation of BUGS code from our model design tools, interfaces to help task authors create tasks from libraries of task- and evidence model templates, and procedures for interacting with experts at both the levels of creating schemas and supplying information about individual tasks created within those schemas.

REFERENCES

Adams, R., Wilson, M. R., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. Applied Psychological Measurement, 21, 1-23.

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement, 23*, 223-238.

American Association for the Advancement of Science (1993). Benchmarks for science literacy. New York, Oxford University Press.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*, 153-168.

Embretson, S. E. (Ed.) (1985). *Test design: Developments in psychology and psychometrics*. Orlando, FL: Academic Press.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*, 380-396.

Formann, A. K. (1985), Constrained latent class models: Theory and applications. *British Journal of Mathematical and Statistical Psychology, 38*, 87-111.

Frase, L. T., Chudorow, M., Almond, R. G., Burstein, J., Kukich, K., Mislevy, R. J., et al. (2003). Technology and assessment. In H. F. O'Neil & R. Perez (Eds.), *Technology applications in assessment: A learning view* (pp. 213-244). Mahwah, NJ: Erlbaum.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman & Hall.

Gibbons, R. D., & Hedeker, R. (1992). Full-Information item bi-factor analysis. *Psychometrika, 57*, 423-436.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. Boca Raton: Chapman & Hall/CRC.

Haertel, E. H. (1984). An application of latent class models to assessment data. *Applied Psychological Measurement, 8*, 333-346.

Hansen, E. G., Forer, D. C., & Lee, M. J. (2001, April). *Technology in educational testing for people with disabilities*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Jensen, F. V. (1996). *An introduction to Bayesian networks*. New York: Springer-Verlag.

Ludlow, L. H. (1986). Graphical analysis of item response theory residuals. *Applied Psychological Measurement, 10*, 217-230.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Mislevy, R. J., Almond, R. G., Yan, D., & Steinberg, L. S. (1999). Bayes nets in educational assessment: Where do the numbers come from? In K. B. Laskey & H. Prade (Eds.), *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* (pp. 437-446). San Francisco: Morgan Kaufmann.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the roles of task model variables in assessment design. In S. Irvine & P. Kyllonen (Eds.), *Item generation for test development* (pp. 97-128). Mahwah, NJ: Erlbaum.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (in press). On the structure of educational assessments. *Measurement: Interdisciplinary research and commentary.*

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (in press). Making sense of data from complex assessment. *Applied Measurement in Education.*

National Reseach Council (2001). *Knowing what students know: The science and design of educational assessment.* Committee on the Foundations of Assessment, J. Pellegrino, R. Glaser, & N. Chudowsky (Eds.). Washington, DC: National Academy Press.

Reckase, M. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9,* 401-412.

Samejima, F. (1969) Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph No. 17, 34,* (4, Pt. 2).

Spiegelhalter, D. J., Thomas, A., Best, N.G., and Gilks, W.R. (1995) BUGS: Bayesian inference Using Gibbs Sampling (Version 0.5) [Computer Software]. Cambridge: MRC Biostatistics Unit. Retrieved from http://www.mrc-bsu.cam.ac.uk/bugs/

Thurstone, L. L. (1947). *Multiple-factor analysis.* Chicago: University of Chicago Press.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185-201.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").